Forum Series on the Role of Institutions in Promoting Economic Growth

Learning *from* Doing:
A Methodology for Self-Evaluating Projects

OMAR AZFAR AND CLIFFORD ZINNES

Forum 5
NIE-Based Toolkits for USAID Applications

Session 3

14 February 2003
Washington, D.C.

Forum Series on the Role of Institutions in Promoting Growth
Directed by The IRIS Center
Sponsored by USAID, EGAT/EM
SEGIR/LIR PCE-I-00-97-00042-00, TO 07

About the Series

The objectives of the Forum Series are to help USAID make its donor assistance more effective and sustainable by incorporating insights from the New Institutional Economics into USAID's programming and delivery of development assistance. Services for the Forum Series are provided by the Center for Institutional Reform and the Informal Sector (IRIS) and its consultants. Editor for the Series is its project director, Clifford Zinnes, with support from the Forums Steering Committee (Ed Connerley, Jim Elliott, Jonathan Sleeper, Tham Troung, and Jolyne Sanjak), chaired by the activity's COTR, Fred Witthans. Funding for the Series is provided by USAID's Bureau for Economic Growth, Agriculture, and Trade, Office of Emerging Markets through the SEGIR/LIR contract PCE-00-97-00042-00, Task Order 07.

The views and interpretations represented in this paper belong solely to its authors and should not be attributed to USAID or to IRIS.

For information, contact:
Dr. Clifford F. Zinnes
Director of Research Coordination
The IRIS Center at the University of Maryland
2105 Morrill Hall
College Park, Maryland 20742
Voice: 301-405-3064
Fax: 301-405-3020
zinnes@iris.econ.umd.edu

Forum 5: NIE-Based Toolkits for USAID Applications

Learning *from* Doing:
A Methodology for Self-Evaluating Projects

OMAR AZFAR AND CLIFFORD ZINNES

The IRIS Center
Department of Economics, University of Maryland University at College Park

February 14, 2003

Executive Summary

There is little systematic evidence or record keeping on the effectiveness of USAID projects (Espina and Zinnes 2003). Moreover, to the extent a project is retroactively assessed, there is no attempt to establish a credible counterfactual, i.e., what would have happened without the project. As a consequence, USAID doesn't know what works and, thus, cannot answer criticisms about the ineffectiveness of its projects.

This leads to a potential "double whammy" for the welfare of the recipients of USAID programs. Funds may be under-allocated to international development and the allocated funds are not used for the most effective projects.

While research using existing data is useful in understanding the nature of this pathology and in informing the design of responses, it only helps us to make educated guesses about what works. To be able to state with any confidence that a reform package works, we would have to "road test" with the appropriate controls. This is the purpose of PREP (Prospective Randomized Evaluation Procedure), the diagnostic toolkit introduced in this paper.

A PREP application involves implementing a reform in some places (the unit of observation could be classrooms, schools or even local government units) and not others. The places which receive the reform are selected randomly from the full sample. The reform is then preceded by baseline data collection and followed by another round of data collection. A comparison of outcomes in the treatment and control group allows us to infer if the reform was effective.

For example, the following reform could be tried to "improve public officials' knowledge of citizens' preferences and match the supply of public services to demand": Select 100 municipalities for the study. Select 50 randomly for the treatment group and 50 for the control group. Ask public officials in all 100 about their knowledge of citizens' preferences. Survey citizens in all 100 municipalities to infer information on preferences. In the treatment group inform public officials of citizens' preferences. One year later, examine whether public officials in the treatment group have (1) better knowledge of preferences, (2) shifted expenditures to the

more demanded public services, and (3) improved the quality of the most demanded public services.

Doing a PREP application is not easy. There are several pitfalls to watch out for. For example, reforms must actually be implemented in the treatment group, and not in the control group. It may be difficult, however, to prevent people in the different groups from knowing what the other group is doing so there could be contamination in the sense that some in the control group become exposed to the treatment.

As an illustration of PREP, we conduct an application for the case of how to evaluate the effectiveness of USAID-funded training workshops for small and medium-scale enterprise (SME) entrepreneurs. In principle, we could have chosen to illustrate PREP for an institutional reform, but time and funding constraints made this infeasible. We allocated participating SME entrepreneurs to 24, simultaneous training sessions in the Philippines and gave the trainers incentives based on the average score of their students on standardized tests. We then examined outcomes based on participant satisfaction ratings of the trainer.

Our results suggest that incentives based on broad outcomes are more effective than incentives based on narrow outcomes for trainers of average or above average ability. However, all incentives might improve outcomes for the less capable trainers. This corresponds with concerns expressed in the theoretical literature on incentives and concerns expressed in the literature on education. There are also several lessons for future PREP applications.

1: PREP applications need to be done on a big enough scale to be sure about the results. In the case of SME training, this means both that more groups are needed for comparison, and that there should be one training shared by all participants so we can anchor their evaluations.

2: There should be a "mini-PREP" before the actual PREP to fine-tune the evaluation technology for a particular application.

3: Policy makers should be careful in recommending pecuniary incentives based on outcomes, especially if the latter are based on narrowly defined measures of outcomes. While these can work, they can also be counterproductive.

4: The additional cost of "PREPping" technical assistance, at least in our sample application was quite small—8 percent—and, as a fixed cost, the percent would decline as the scale of the technical assistance rises.

Forum 5: NIE-Based Toolkits for USAID Applications

Learning *from* Doing:
A Methodology for Self-Evaluating Projects

OMAR AZFAR AND CLIFFORD ZINNES

14 February 2003

1 Introduction[1]

It is increasingly clear that institutional failures are the primary cause of under-development. However, the statistical study of the determinants of good institutions or good governance is limited by the fact that institutions vary mostly across countries, and cross-country regressions suffer from problems caused by omitted variable bias, simultaneity bias and selection bias. Also there may simply not be enough institutional variation across countries to examine some interesting hypotheses. Prospective reformers therefore find insufficient guidance in the existing econometrics literature on which reforms would be most effective. Occasionally evidence is available at the micro level for institutional reforms that work, but such evidence often offers little systematic guidance. Sitting at a recent high-level meeting on anti-corruption reforms, one of us was struck with how close our state of knowledge was to medieval medicine. We were aware of possible impacts in isolated incidents of certain reforms, but did not really know if those reforms had had the stated impact on corruption, or even if the level of corruption really had moved or not.

There is little question that causal relationships in terms of governance and socio-economic outcomes are complex and difficult to disentangle. But such causal relationships are also difficult to disentangle in the medical sciences. Yet medical science has made tremendous progress in the past century in the design of responses to pathologies. Life expectancy in most

[1] We are deeply grateful to Jonathan Alevy, Dan Blumhagen, and Peter Murrell for their insightful commentary on an earlier version of this paper. Their collective contributions have helped to make this paper more closely aligned to USAID's needs and the real-world conditions facing experimental economists. Any errors remain wholly our own.

countries in the world by the early 1980s had reached the levels of the richest countries at the turn of the century. This was in remarkable contrast to the non-convergence in incomes. What was the secret of their success? Many of the most important discoveries of modern medicine were in fact accidental, and based on poorly understood phenomenon. But modern medicine had adopted the idea of the randomized clinical trial (Kremer 2002). This procedure allowed the examination of the effectiveness of treatments *even without* a complete understanding of the causes of ailments. A prominent scientist proclaimed that the most important advance in modern medicine was not any one discovery but rather the randomized clinical trial. Similarly, effective reforms can be designed and identified with imperfect knowledge of the causes of institutional failure.

In modern medicine, lessons are drawn from complex real world phenomena and exploratory lab work, but these lessons are primarily used to educate the guesses of scientists who then design rigorous experiments to test what works. In the first instance, researchers conduct these experiments on laboratory animals but eventually conduct them on humans before being permitted to try new treatments on the general public. Similarly, policy reforms should first be tried in classroom experiments—where they can be tried at low cost—and later on actual political institutions before being broadly recommended.

In recent years, classroom experimental methods have entered the mainstream of economic analysis, with experimental papers now appearing regularly in all the major journals. Indeed experimental economics has forced economists to rethink their cherished assumptions of rationality, and led to a *gestalt shift* in economic thought. This year the Nobel prize in economics was awarded to two scientists, Daniel Kahneman and Vernon Smith for their work in experimental economics.

Such experimental data suffers from a lack of contextual validity.[2] Laboratory incentives imperfectly mimic real world situations and the stakes are usually much lower. Thus while

[2] Such experiments are typically conducted using undergraduates at American universities, primarily for convenience. For policy reforms aimed at developing countries it may make sense to conduct experiments in these countries, with the subjects drawn from the relevant populations. This would be more costly in terms of researchers' time and inconvenience but would probably allow lessons to be drawn more credibly for the country in question. Running experiments abroad has the additional advantage that the stakes of around $30 typically used for American undergraduates would represent important choices for subjects in countries where this may represent a month's income. Subjects could thus be more reasonably expected to act in experiments as they would when facing important real life choices. Still stakes would generally not be as high as important real life choices.

useful, a "classroom experimental analysis" may not provide sufficient evidence for a widespread policy reform. It may however provide enough evidence for a "real policy experiment". In fact real policy experiments have recently been tried, for example with schools in several developing countries (Kremer 2002).

This paper is structured as follows. We begin with a discussion of how a "Prospective, Randomized Evaluation Procedure" (PREP) can be used to evaluate institutional reforms. Here, we take the opportunity offered by the widespread devolution of authority that has given meaningful responsibilities to hundreds of local government units in some countries. We discuss how reforms can be tried in some units and not in others in a randomized manner, which then would allow an evaluation by comparing the treatment and control groups. The section discusses how we might make randomization politically palatable, as well as how we would address other problems with randomized evaluation.

We then describe a PREP application that we conducted in the Philippines. In principle, we could have chosen to illustrate PREP for an institutional reform in dozens of municipalities, but time and funding constraints made this infeasible. Instead, as an illustration of the technique we conducted an application to show how one might design and implement a PREP evaluation into a training program. We selected the simple case of marketing workshops for SME entrepreneurs. The central question was the effectiveness of providing incentives for trainers. Incentives are presumed to have an effect on behavior but there are several concerns about the effect of incentives on teachers and trainers, or for that matter any activity where outcomes of real value are difficult to measure (Prendergast 1999; Dixit 2001). To address these concerns we offered trainers incentives based on different outcomes and compared their performance across groups. Our results suggest that, as theory predicts, incentives do have an effect on outcomes but, if inappropriately designed, can be ineffective or even counterproductive. Our results resonate with the concurrent work done by Glewwe *et al*. (2003) who find that in Kenya the provision of monetary incentives to teachers only led to short-term improvements in student performance.

Finally we offer recommendations on several questions. First, with the benefit of hindsight, we point out how better to conduct a study like this one so as to obtain clearer results

in the future. Second, we suggest how to structure incentives for trainers in The Philippines. Third, we show how to design USAID programs to allow for rigorous evaluation.

2 PREP applications to anti-corruption and other local government policy reforms

Let us now illustrate the use of PREP by considering a number of policy applications before getting down to the detailed definitions and structure of a randomized prospective evaluation in the following section. We start with the case of decentralization and then briefly turn to other potential reforms related to local government.

2.1 Decentralization and the opportunity for randomized evaluation of policy reforms

The recent wave of democratic reform has brought in its wake the devolution of power from central to local governments across the world. Several countries like Indonesia, India, Brazil, The Philippines and Russia have devolved meaningful responsibilities to hundreds of local government units. The widespread decentralization of authority in many countries offers the possibility of conducting policy reform in a randomized manner. By "randomized", we mean that reforms can be conducted in some municipalities and not in others, and outcomes can be compared in the control and treatment groups. Because many countries have hundreds of local government units with meaningful responsibilities, there is reasonable hope that varying characteristics of the reform or its implementation might lead to finding a significant impact for some variations of the reform and therefore whether such reform programs are actually effective. Conducting randomized evaluations, and eventually creating the capacity to do such evaluations of reforms in developing countries, will be an important step in helping developing countries learn about which reforms actually work.

Example of a PREP study on institutional reform. Because institutional reforms are difficult to implement and many reforms are likely to fail, either in implementation or in impact, an evaluation design for institutional reform should try several of them at once to increase the probability that at least one would be identified as effective. The ideal experimental design may involve 200 municipalities being chosen for a study of 3 different reforms (say) community oversight, accounting reform, and wage reform. Forty municipalities would get no reform, forty would only get oversight reform, forty would only get accounting reform and forty would get only wage reform. The remaining forty municipalities would get all three reforms because it is possible that there are complementarities between reforms and some reforms only work when

others are in place.[3] For instance accounting reform that made it easier to identify corrupt politicians may become more effective if community oversight were increased.

Data would be collected on corruption before the reforms were implemented in all 200 municipalities using both survey methods and at least one non-survey method. Next, data would be collected on whether reforms were actually implemented.[4] Finally, data could be collected a year or so after the reforms were implemented in all 200 municipalities. A comparison of changes in the quality of service delivery or the level of corruption in the different groups, along with an evaluation of whether reforms were implemented, would allow us to determine which reforms work. Table 1 summarizes the steps in such a study.

Table 1: Steps and time table for a PREP application to local government policy reform

Step	*Action*	*Timing*
1	Design anti-corruption reform with a randomization protocol and "sell" randomized reforms to recipient country.	Months 1-4
2	Design baseline measurement of outcomes and run informativeness tests on them.	Months 1-4
3	Implement reforms according to randomization protocol.	Months 5-6
4	Check that reforms were de facto implemented by collecting data from public officials.	Months 8-9
5	If reforms were only effectively adopted in some places, examine what predicts whether reforms were adopted or not.	Months 10-11
6	Conduct follow-up survey a year after reform.	Months 18-19
7	Compare outcomes in control and treatment groups	Months 21-24

We now turn to a discussion of the various kinds of institutional reforms that might be tried. We take the examples of four possible reforms that are complementary to decentralization: "improving knowledge of preferences", "sharpening inter-jurisdictional competition", "improving accounting standards", and "raising wages".

Improving knowledge of public's preferences. One of the arguments for decentralization is that local government officials know better the preferences of the local population and can better match public service provision to the demands for specific services. However, public

[3] In principle, all possible permutations of reforms could be tried but this would increase the number of cells and would either raise costs substantially or reduce the number of observations in each cell and jeopardize the possibility of finding statistically significant results.

[4] This could have interesting correlations with the information collected in the baseline study. For instance it would be worth examining whether initial levels of corruption, especially elite corruption are correlated with non-adoption of reform.

officials may have only weak knowledge of the preferences of citizens (Azfar, Kahkonen and Meagher 2001). It is possible that simple informal interactions between public officials and the public will allow for this, but many of these sub-national units are quite large, with populations of millions of people, and local officials are more likely to socialize with the elite than the general population. Therefore, informal information flows may be quite imperfect modes for conveying information on the population's preferences. Thus, reforms that improve information flows on preferences can have a positive impact on matching supply to demand and, therefore, on the value of publicly provided services.

One straightforward way of improving officials' knowledge of citizen preferences is to collect data from the public on their preferences for various public services (using surveys) and to convey this information to public officials with authority over resource allocation. This information could also be published in the local newspapers so the citizenry can discipline the local government to provide the services demanded. Another way to think about this question is that local government officials have all kinds of reasons to distort public expenditure priorities to benefit either themselves personally or their friends in the private sector. The absence of systematic information on what the public wants makes it easier to do this in the cloak of public interest. A systematic collection and dissemination of information on the public's preferences would make it more difficult to distort public expenditure priorities.

Implementation can be done by surveying citizens and firms about their preferences for publicly provided goods. Respondents would have little incentive to lie, in fact, they have an incentive to tell the truth as resources might be focused on services they say they would like provided (though admittedly there may be some strategic voting issues).[5] Several variables can be used for evaluation. First, we can gauge whether public officials improve their knowledge of public preferences after these preferences are communicated to them—they might simply never pay attention to the information or forget it. Second, we could examine the local government's accounts to see if resource allocations actually changed in response to the information communicated. Third, we could examine citizens' satisfaction ratings of various public services—did

[5] Like any other variable, collecting information of preferences is not trivial. Respondents can most easily be asked about their most preferred public service, but this only provides information on their most preferred public service and not on other highly ranked services which may also be important. While in principle they can be asked to rank several public services, they are liable to get confused and provide erroneous answers. In our experience, we have been able to elicit information on as many as three most desired public services.

satisfaction ratings improve for public services on which citizens said they wanted public expenditures focused? Fourth, we could do independent evaluations of the improvements in service delivery—did roads actually get built, did water supply actually become more sanitary etc.?

Sharpening competition between local governments. According to the theory of fiscal federalism, one important argument for local government is that it allows governments to compete against one another to attract income-generating—and therefore potentially tax-paying—households and firms. This desire to expand the tax base might induce the local government to provide better services. However, there are also concerns about a "race to the bottom" by such a competition.[6] What happens, therefore, can be studied by setting up such a competition and observing whether sharpened competition leads to better service delivery, or increases in cross-jurisdictional inequality (or possibly both).

In the case of "sharpening competition" the actual reform proposal has several steps. Lists of 10-20 proposed reforms will be discussed with experts, businessmen and citizens. The five reforms that are generally thought to be the most important would then be offered to those municipalities in the treatment group. At least two of these should be reforms that are relatively easy to adopt—this will build the confidence of those municipalities that think reform is an impossibility. Municipalities will then be ranked according to how many of these reforms they adopted and these lists will be made public. Subsequent to this, an evaluation would be undertaken as to whether citizens or firms really do move to areas where more reforms are implemented, and whether there were improvements in public services in the treatment group.[7]

It is important to state that comparisons have to be made between groups which were offered the reform and those that were not. Since reforms would be adopted in a non-random manner within the treatment group, a comparison of municipalities within the treatment group that adopted particular reforms and those that adopted other reforms would not be valid, and would suffer from the usual interpretation problems associated with retrospective, real-world data.

[6] This refers to the phenomenon in which jurisdictions compete for (or lure) business by offering ever-sweeter tax deal (or inefficiently lax regulatory standards). Here, competition serves to lead to under-funded and under-supplied public services, i.e., a race to the bottom.

[7] Here, an improvement is measured more broadly than simply whether these reforms were adopted.

The use of PREP has added advantages here in that it allows one to control for the effects of initial conditions. For example, one could evaluate whether particular forms of corruption like elite capture identified in the baseline survey predicted slow adoption or non-adoption of reforms. This could be done by relating the initial level of corruption to the number (or degree) of reforms the municipality ultimately undertook. In other work one of us has found that initial levels of corruption delay the adoption of trade reform (Lee and Azfar 2000); an application of PREP to decentralization would permit an analogous examination of the effect of initial levels of corruption on subsequent reforms adopted by local government units.

Increasing Transparency. Transparency can be thought of as increasing the probability of exposure of corrupt or incompetent officials. Examples of increasing transparency are standardizing accounting systems and requiring public officials to regularly declare their assets. These changes would make it easier for law enforcement officers to expose corruption. Several examples are reported by Klitgaard's (1988) classic, *Controlling Corruption* where he describes the successful anti-corruption programs in Hong Kong, Singapore, and The Philippines. In all three cases, an improbable amount of wealth for a public official was used as the basis for initiating investigations. In Hong Kong and Singapore the law was amended to allow for disciplinary action on the basis of unexplained riches, i.e., in the presence of unexplained riches the burden of proof was transferred onto the accused. Random and surprise wealth checks were also used to good effect in all three countries. Klitgaard also reports that many of the successful anti-corruption policies that were adopted in Hong Kong, Singapore, The Philippines and Bolivia can be thought of as increasing the probability of exposure. These include improving accounting and audit systems, checking on bank accounts, hiring undercover agents, using reports from the media and public, and providing incentives for officials to report the offer of bribes.

Another example of the importance of transparency is the Ugandan experiment on public expenditure tracking. Clear declarations by the government regarding how much money was allocated to each delivery point may have reduced leakages from the government budget.

The reforms mentioned above could be legislated and implemented in some municipalities and not in others. A comparison of outcomes in the treatment and control groups would demonstrate the effectiveness of these reforms. It is important, however, that reforms not merely be evaluated by outcomes closely related to the reform. For instance, an evaluation of expenditure tracking should involve more than evaluating a reduction in leakages as reported by the

tracked accounts. These are likely to change but this doesn't necessarily imply that actual leakages have fallen. Effects on actual outcomes must be demonstrated against a counterfactual or control group to make a persuasive case that expenditure tracking works in reducing corruption and improving service delivery. Similarly, outcomes directly related to the public declaration of wealth of public officials (like changes in their apparent wealth) cannot in themselves offer reliable evidence of a reduction in corruption. Public officials may just have become cleverer about hiding their wealth (passing it on to their relatives, moving it offshore, etc.). Other evidence of reductions in corruption must be sought out.

Increasing Government Wages. Increasing wages should have two effects on the level of corruption. First, there is an income effect. A wealthy official has less to gain in utility terms from being corrupt. Second, and probably more importantly, there is an incentive effect. The possible loss of a relatively well-paying executive position may limit corruption. Again, Klitgaard's accounts contain several real-world examples of the importance of wage reform in an anti-corruption strategy. The successful anti-corruption efforts in Hong Kong, Singapore, The Philippines, and Bolivia, included both increasing the pay as well as making promotions and retirement benefits contingent on good, honest work (Klitgaard 1988, Klitgaard *et al*. 2001).

There is a continuing debate on the importance of wage reform in reducing corruption. Some anecdotal accounts suggest that government wages affect corruption. A devaluation that dramatically lowered real wages of government officials in Cameroon was reportedly followed by a sharp increase in corruption. In our personal experience discussions regarding corruption in Pakistan often highlight the impossibility of living on a civil service salary. However, statistical evidence on the link between wages and corruption appears mixed. While van Rijkehem and Weder (2001) find that government wages are related to lower levels of corruption, others (Lederman, Loyaza and Soares 2002; Rauch and Evans 2000) state that this result is not robust.

Even if the partial correlations between government wages and corruption are statistically significant, they may not allow any clear inferences, because government wages may be correlated with other measures of good governance—like the ability to collect taxes or keep public employment at sensible levels. Therefore any correlation between government wages and corruption may be spurious or driven by reverse causality.

There is another problem with inferring a relationship between wage reform and corruption based on a cross-sectional examination of wages with corruption. It is possible that while

high wages do prevent the emergence of corruption, once corruption has emerged and the practice of the sale of jobs has become entrenched, increases in wages would simply be capitalized into the price of the job, and wage reform would not reduce corruption (unless accompanied by a substantial increase in the probability of being fired for being corrupt).

Table 2: How to implement and evaluate local government reform

Reform	*How to implement*	*How to evaluate*
Improving public official's knowledge of citizen's preferences	Collect data on preferences and communicate results to officials. (Since researchers can do this on their own or with their own team checking implementation is an internal process)	✓ Ask officials about citizen's preferences several weeks after they got the results to see if they forgot (or perhaps never learnt) about citizen's preferences ✓ Examine public accounts several months after report to see if funds got reallocated to demanded categories ✓ Survey users to see if service delivery on the most demanded services improved
Sharpening competition between local government	✓ Design 10 reforms, hold focus groups and desk studies and chose 5 to implement ✓ Offer localities in treatment group the chance to enter a "5-star" competition based on adopting these reforms	✓ Examine if the users of public services report larger improvements in the treatment group than the control group. ✓ Examine if there is net migration into the treatment group.
Increasing transparency in local governments	Legislate better accounting standards, requirements for public declarations of assets, or expenditure tracking systems for the treatment group and check compliance with them.	✓ Examine if leakages from public funds have declined in the treatment group ✓ Examine if there are increases in the prosecution of officers in the treatment group ✓ Examine if citizens and firms report less corruption ✓ Examine if public services have improved
Increase government wages	Increase wages in treatment municipalities, and check public officials are actually being paid more.	✓ Examine if the "price of jobs" has increased ✓ Examine if public officials report that corruption has declined ✓ Examine if firms and households report that corruption has declined

Results from randomized evaluations can be free of these concerns and can examine whether higher wages lead to lower corruption. Evaluation of these reforms should be done by surveys of both service users and public officials. (Public officials would have an incentive to say wage increases reduced corruption because they would like more wage increases). Also some non-survey method should also be used to detect the reduction in corruption (see Azfar and Murrell 2003).

2.2 Other local government reforms

There are several other reforms which are amenable to PREP and which can plausibly have an impact in terms of improving service delivery or reducing corruption in some environments. These are summarized together with the PREP steps in Table 2.

Public officials could be rewarded or punished on the basis of performance. While not explicitly an anti corruption reform, such a reform may have larger impacts on the level of corruption than reforms that try to directly target corruption. For instance sampling allegedly immunized children, checking the saliva for antigens and punishing health care workers for the absence of antigens may be more effective than a reform targeted at the illegal sale of vaccines to private providers.

Privatization or "NGOization" of service delivery. It is often stated that privatization or "NGOization" would reduce corruption but this is seldom rigorously evaluated. Private providers and NGOs can also siphon off or waste funds and perform poorly in terms of service delivery. Privatizations and NGOizations could be conducted in some municipalities before others and impacts on corruption and service delivery could be studied.

Transparency in public procurement. This might include requiring detailed documentations of public procurements, how bids were solicited, which ones were received, and the basis on which the award was made. The compulsion to produce such documentation would make it more difficult to award the bids to poor options. One outcome variable that could be used is the quality of the documentation which is often incomplete or poor in developing countries. Another could be patterns in the bids which suggest collusion amongst bidders.

Implementing an IT system to track multiple prescriptions, shadow patients, etc. Such systems are used to uncover white collar crimes in developed countries. If compliance with accounting standards could be insisted upon, similar systems could be used to track public sector accounts in developing countries.

Increasing community oversight of public services. This has been tried in several places, and results from El Salvador and Bolivia suggest that such oversight does improve service delivery.

3 The ingredients for a successful "prospective random evaluation"

Let us now highlight the key ingredients as well as the structure of a successful prospective, randomized evaluation. It is instructive to clarify what is required for making inferences about a broad reform package from a prospective randomized evaluation. These are listed in Table 3. (At the end of the paper we will indicate the scope for relaxing these requirements.)

Table 3: Ideal conditions for drawing an inference from prospective randomization

1. Outcomes must be measurable
2. Implementation must be measurable
3. Reform must be implemented according to a clear randomization protocol
4. There must be a sufficiently large number of "individuals" in control and treatment groups
5. Treatment group must adopt the reform just as they would if the reform were broadly implemented
6. Control group must continue to act as it would if the reform were not implemented
7. There must be no spillovers in outcomes from the treatment to the control group or vice versa
8. Outcomes must be measured before reform and long enough after reform for the impacts to be apparent

Let us now highlight three important ingredients in a "real" policy experiment. The first is that we must be able to measure impact. Here we recommend that several different impact measures be used so that it is possible to run robustness checks on the results and increase confidence in them. The most informative of these can eventually be used for project evaluation. The second is that the implementation of the reform program be randomized, that is, that there be a control groups and that variations of the reform be randomly applied to the treatment groups. This is politically non-trivial in real-world settings but good arguments can be made to allow this to be done. The third is that the reform actually be implemented, i.e., the reform must be more than *pro forma*. We refer to these three steps as "Evaluation", "Randomization", and "Implementation". We now discuss these in turn.

Evaluation. Evaluating a reform requires the measurement of governance or other relevant outcomes both before and after the reform. Reports from households or firms on public services can provide reasonable information. However, it probably makes sense to cross-check these reports with other kinds of data. For instance, it may make sense to compare household and firm reports on electricity outages with data collected directly from observing the electricity

supply at the regional offices of the survey firm, or the houses of enumerators. Of course one doesn't expect a perfect match but the two kinds of data should be correlated.

Measuring the impact of anti-corruption programs on corruption itself very difficult. Corruption is nowadays typically measured by surveys but there are serious concerns about the quality of this information. IRIS is currently conducting a project which will help evaluate the accuracy of survey responses but this in itself is a tricky exercise.[8] Our take on the use of surveys to measure corruption is that while they are useful, there are enough concerns about the quality of the data that they should be used in conjunction with other variables when evaluating corruption.[9]

Once data has been collected it should be checked for "informativeness". Essentially this means two things. First we should expect that evaluations by different people in the same "group" (e.g., local government unit, school, class-room, etc.) should be correlated. Formally, this can be examined by looking at whether there are significant variations across "groups". Second, if different variables are purportedly measuring the same underlying characteristic, they too should be correlated.

Sometimes evaluators are fortunate enough to have information on a variable, for which such a strong presumptive case can be made for its effect on outcomes, that a "good" performance variable can be expected to be correlated with it. In this case a simple correlation of performance variables with this variable can provide a reasonable validity test (this is what we do in section 4.2).

Randomization. The second important component is randomization of program implementation. At first this seems impolitic, but with some consideration it is possible to understand how most people would benefit from such an exercise. Let's take an example of a country like Indonesia, The Philippines, India, Brazil or Russia where there are a large number of sub-national governments. The main reason we are planning to try reforms as an experiment is that we have made some educated guesses that they might work, but have no clear knowledge of the effectiveness of reforms. If by some chance there are municipalities where we are so convinced the reform would work that it would be immoral to deny them the reform, then we can

[8] This work is described in Zinnes and Azfar (2003).
[9] See Azfar and Murrell (2003) for suggestions on how else to measure corruption.

implement the reforms in those municipalities and select them out of our "study sample". The same holds for municipalities where we are convinced the reforms will not work. Amongst the rest we can randomly assign and run alternative versions of the reform in participating municipalities. Because there are hundreds of local government units with meaningful responsibilities in these countries, a large enough sample could be selected so that, if the effects of the some versions of the reforms were of reasonable magnitude, they would be detected.

Two arguments can be made to make randomization politically feasible. First, since randomization, if explicitly done, is almost by definition fair, the government or donor cannot be accused of playing favorites. (Here, such a study should be "sold" before the random selection is made to increase political support for it). Second, those municipalities not selected in the first wave of reforms will have the benefit of experience and receive more effective reforms for their patience. Thus, a prospective, randomized implementation of an anti-corruption reform can be "sold" politically with some thought and packaging. (In a sense, it is not clear why there is so much concern about the impoliticness of randomized implementation—after all, "pilot projects" are conducted all the time, only that they are just not done in enough places or with enough care about collecting data on the counterfactual to be generalizable).

Implementation. The third requirement for an appropriate policy experiment for evaluating a reform is making sure it was in fact implemented. Thus after design and implementation, surveys of providers and public officials should be carried out to ensure the reform was not merely *pro forma*. (Such a confirmation is, in a sense, a kind of preliminary evaluation). Take the example of increases in the transparency of accounting systems or increases in government wages. A government may change the rules about accounting standards but not actually implement these rules. To make a clear statement that "changes in accounting standards do not reduce corruption" one would have to make sure that accounting practices had changed and corruption had not. Wage reform, too, may simply not be carried through in practice, especially when budgets are tight. Examining the effect of wage reform would first have to establish whether wage reform was implemented and then whether corruption levels changed.

It is also worth checking that reforms have not been adopted in the control group. Municipalities may learn from their neighbors and adopt popular programs even if they are not formally part of the treatment group. Since these reforms may be adopted in the municipalities where they are most effective, this may confound the results even if only a minority of municipalities in

the control group adopts the reforms. Furthermore, human beings do not always like being experimented on and may change their behavior, confounding experimental results. There is also the more basic issue of the effect of observation on behavior. The prospective random evaluator must be savvy about all these issues, and design a protocol for detecting each possible problem in group comparability.

4 An application of PREP: Evaluating the impact of trainer incentives

To further illustrate exactly how to use PREP, we now describe an example of an actual application of a prospective evaluation procedure we conducted in The Philippines as a part of the Forums Project. Due to time and budgetary limitations, we selected a more manageable problem than institutional reform of local governments. Instead, we chose to apply our methodology to the issue of how to make training programs, which USAID frequently funds, more effective.

As a concrete case, we picked the training of SME managers in strategic marketing. Our evaluation was focused on the impact of incentives on outcomes. We offered different groups of trainers different incentives and compared outcomes across groups.

4.1 Theoretical background and implementation design

In order to fully understand this application, let us provide the background for this exercise in terms of the theory of incentives, and the relevance of these ideas for USAID.

Theory and relevance of performance-based incentives. That monetary incentives matter and have an effect on behavior follows easily from the assumptions of rational choice and utility maximization. The early formal work on the theory of incentives showed that incentives could improve performance and empirical work on the provision of incentives in firms showed that incentives did improve performance.[10] However, what these theoretical and empirical studies had in common was that the "principal" could observe something closely related to the principal's interest—the funder of the trainers and the quality of the latter's teaching in the present example. Changing the premise of observability can call the value of incentives into question.

[10] See Prendergast (2000) for a survey.

Subsequent theoretical work has in fact shown that if the principal cannot observe whether his objectives are being maximized, then incentives can be distracting rather than motivating (Holmstrom and Milgrom 1991). The issue has also been highlighted in policy debates on subjects like education where observable variables like scores on standardized tests are thought by many experts to distract rather than motivate teachers. Teachers may "teach for the test" in response to such incentives, they may encourage students to concentrate on doing well on items the teachers' incentives are based on and, in extreme cases, may even lead teachers to turn a blind eye toward student cheating.

These insights are clearly relevant for USAID since, in general, the final outcomes of aid programs are not easily measurable and explicit incentives may have to be based on intermediate results. This may lead USAID's implementers to improve performance as measured by these intermediate results without achieving genuine improvements in performance. Thus, the concerns we raise in the discussion and results presented in this paper have a broader relevance to aid programs.

The theory of incentives also suggests that incentives based on broader measures of performance are more likely to motivate "agents" toward the objectives of their principal than incentives based on narrower measures of performance. To test this theory we offered trainers incentives based on outcomes of different breadth. How this was done in described in the next section.

Experimental design for PREP of incentives for trainers in The Philippines. The trainings were conducted by the Institute for Small Scale Industries (ISSI) at the University of The Philippines. ISSI routinely conducts trainings for SME managers and proved to be a good partner. After consultations with ISSI and USAID, we decided to pick strategic marketing as the subject to be taught.

The training was done in two stages. ISSI hired Dr. Felix Lao, the author of a widely used textbook on strategic marketing in The Philippines, to train the trainers. On 10 November 2002, Dr. Felix Lao trained 30 trainers who were subsequently to train SME managers. All trainers took an exam at the end of the "trainer training". Over the subsequent week, trainers were asked to give trial presentations of how they would teach and their performance as teachers

was graded by ISSI staff.[11] We would later use the exam score and trial presentation grade as two controls when assessing incentive effects on performance.

Though we began with thirty trainers, one trainer left during the training and another did not show up for the trial training. Both were taken out of the sample. Of the remaining 28 trainers, we removed the one who got the lowest score, leaving us with 27 trainers.

The 27 trainers were randomly placed in incentive groups according to Table 4. Eighteen trainers—one per group—were provided incentives based on the percentage of responses the students in their group answered correctly on an exam administered at the end of the training. For these eighteen groups, six groups had exams based on 20 questions, six groups had exams based on 40 questions, and six groups had exams based on 80 questions. The possible monetary values of these incentives are listed in Appendix C and vary from 0 to 10,000 Pesos (approximately 0 to 200 U.S. dollars). Eight trainers did not have incentives based on their students' performance. These trainers got a fixed, extra payment of 3,000 Pesos, which we guessed would be the average amount of the incentive payment.[12] Information on their incentives and the questions their students had to answer were sent to them by mail a week before the training. One failed to show up on the day of the training (24 November), reducing Group 1 by one trainer.

Table 4: Randomization matrix for PEP Training Study

Incentives	*Number of questions on test*	*Number of groups*
No	20	2*
No	40	3
No	80	3
Yes	20	6
Yes	40	6
Yes	80	6

* In fact, 3 had been planned, but one trainer did not show up.

A total of 274 participants comprising SME owners or senior managers showed up for the training. They were randomly placed in classes of approximately equal size. All participants provided some basic information and took a pretest of 20 questions. Students then had a five-hour training on strategic marketing (with a lunch break). This was followed by a multiple-

[11] A set of specific criteria, each graded from 1 to 5, was used to minimize the subjectivity of these teaching performance evaluations.

[12] In the event, it turned out we underestimated the average incentive payment by around 25 percent.

choice test of varying length. Students got time proportional to the number of questions they had to answer. This was followed by two essay questions. Finally, we asked several questions about their rating of the trainer as a teacher, whether the training would change how they conducted their business, and whether they would be willing to pay for subsequent trainings. The data was then computerized by Asia Research Organization (ARO) and checked by ISSI and emailed to the authors for analysis.[13]

4.2 Measures of outcomes

As alluded to in section 3 the success of an evaluation depends on the construction of the appropriate indicators of outcomes and performance. For the present application, we used several different indicators to measure outcomes:

- Scores on multiple choice tests;
- Scores on essay questions;
- Scores on satisfaction ratings of the trainer;
- Scores on whether the training changed the way the students would do business; and
- Scores on the willingness to pay for subsequent trainings.

Let us discuss each of these in turn. (Table A.1 has the means and summary statistics of these performance variables.)

Scores on multiple-choice tests. There was one set of twenty questions answered by all students. Therefore, scores can be compared across groups to produce one kind of evaluation of trainer performance. We would expect to see higher scores in this performance measure for students taught by trainers offered an incentive since the monetary incentives were explicitly based on their students' multiple-choice scores. However, for the purposes of donors and policy makers, this could be an unsatisfactory measure of performance as trainers may simply be "teaching for the test". We therefore also measures performance in several other ways.

Scores on essay questions. Students were asked to answer two essay questions. The questions were then graded by two graders according to specific criteria. We used two graders so we could check for consistent grading: we expected the grades given by the two graders to be correlated. The variable we use for the essay scores is then taken as the average of four constitu-

[13] ARO is a professional survey firm which the authors are using for a companion study (Zinnes and Azfar 2003) on the transaction costs facing businessmen in The Philippines.

ent variables (i.e., each of the two graders' scoring of two questions—2 questions times 2 graders). All four components are significantly correlated with each other as we would expect. The scores on the essay question may offer a better evaluation of teaching because the trainers' incentive payments were not based on the scores of the student's scores on essay questions, and thus higher scores are likely to reflect broad gains in learning and not simply teaching for the test. Also many educators believe that essay questions provide better evaluations than answers to multiple choice questions.

Scores on satisfaction ratings. We asked students a number of questions to determine their satisfaction ratings of their trainer.[14] Many of these questions were similar to satisfaction ratings used in American Universities. We had some initial concern about the relevance of the questions to The Philippines context but found that ISSI actually uses similar tools. (The Philippines is quite Americanized by developing-country standards). There were questions about how clearly the lecturer spoke, whether she was well-prepared, etc. and a general question asking for the overall rating of the lecturer. As expected, all the components were significantly correlated with each other. An aggregate score based on all these ratings is used as one measure of trainer performance.

Scores on whether the training will help change the way firms do business. When supporting programs to train SME managers, USAID is primarily interested in changing the way managers do business. The strategic objectives are to encourage growth and trade. Thus, in addition to acquired knowledge and student satisfaction with trainers, it is also important to ask about whether the training would change the way firms do business. (Of course ideally we would go back to The Philippines and examine whether the training actually *did* change the way firms do business, but this was not possible within our time frame). We, therefore, asked several questions about whether participants believed the training would change the way they would market their products. All the components of this set of questions were significantly correlated with each other as we would expect. An aggregate of these answers is used as another measure of performance.[15]

Willingness to pay for subsequent training. Economics teaches us that we can measure the value of a good or service by the amount people are willing to pay for it. Therefore, we also

[14] These satisfaction rating questions are listed in Appendix B.1.
[15] These questions are given in Appendix B.2

asked questions on the willingness to pay for subsequent trainings by the same trainer as a proxy for the quality of the training. These questions were asked in two different ways.[16]

First, we asked whether the students were willing to pay a *fixed amount* (which we specified) for another training by the trainer on four different, albeit marketing-related topics. We then constructed a first willingness-to-pay indicator by aggregating participant responses within each respective group. Second, we asked what the *maximum* the participant would be willing to pay for another training on the same four topics. Respondents were allowed to choose one of five amounts: 1000, 1500, 2000, 2500, or 3000 Pesos. This allowed us to construct a second willingness-to-pay indicator by aggregating participant responses within each respective group. Summary statistics are provided in Appendix A.1 for the first indicator in rows H11-H14 and for the second indicator in rows H15-H18. As expected, all the components across the rows were significantly correlated with each other.

Table 5: Correlation of performance variable with trainer quality

Performance indicator	*Correlation coefficient and P-value*	*Significant**	*Robust regression t-stat and P- value*	*Significant**
Average multiple-choice score	0.04 0.46	No	-0.62 0.58	No
Average essay score	0.07 0.21	No	1.08 0.28	No
Satisfaction rating	0.13 0.09	**Yes**	1.87 0.06	**Yes**
Training will change business practices	0.11 0.21	No	0.11 0.91	No
Willingness to pay (fixed value)	0.07 0.35	No	Did not converge	No
Willingness to pay (maximum value)	0.01 0.82	No	-0.30 0.76	No

* At the 90-percent level of confidence.

Our experience as students and educators has been that the quality of teaching significantly affects learning and the value of the lesson. An obvious way to check for the validity of performance variables is to correlate the performance variables with the trainer score. The trainers were graded on their answers to the multiple choice test, essay questions, and graded on

[16] These willingness-to-pay questions are listed in Appendix B.3.

teaching ability by ISSI experts. A composite measure of these scores was used as the trainer quality variable. We calculated the raw correlations of the performance variables with the trainer quality variable, and also ran robust regressions to examine if these correlations were robust (Table 5). In each case only the satisfaction rating variable showed a correlation with the trainer quality variable. We also found that the satisfaction ratings were more correlated with components of the trainer score variable than were the other performance variables.[17] For this reason, we only used satisfaction ratings as a performance variable when investigating the determinants of trainer performance. The other performance variables seem to contain so much noise (other random causal factors) that even trainer quality doesn't seem to affect them.[18] This makes it *a priori* unlikely that a variable, like the incentives we offer, would have a perceptible impact on these measures of performance.

4.3 The determinants of trainer performance

We now begin our examination of the determinants of the performance variables that seems to capture the quality of teaching: satisfaction ratings and (perhaps) essay questions. But first let us recap the theoretical predictions.

The more recent theoretical work on incentives has highlighted how incentives based on indicators not closely related to the objectives that matter most can be distracting rather than motivating (Holmstrom and Milgrom 1991).[19] This theoretical prediction has been subject to empirical examination in the field of education. These concerns about incentives for teachers were voiced even before the theoretical advances by Murnane and Cohen (1986). More recently, Eberts, Holleneck and Stone (2000) have shown that a school that introduced merit pay based on student retention, improved its performance as measured by retention rates, but may have undermined performance more broadly defined. However, their paper involves the comparison of only two schools, making any clear inference difficult. In developing country contexts, Glewwe *et al*. (2003) have shown that monetary incentives for teachers in Kenya only led to short-term improvements in performance.

[17] These results are not shown in Table 5 to save space but are available from the authors.
[18] As discussed below, had this level of noise been anticipated, it can be partly corrected by setting a larger sample size. Normally, such problems can be avoided by administering a pre-test to establish *inter alia* the adequate sample size. Unfortunately, time limitations for the application prevented this step from being implemented.
[19] See Dixit (2001) and Azfar (2002) for a review of this literature.

Theory predicts that incentives can improve performance but may also encourage trainers to try to "game the outcome". Trainers may try to teach for the test. Students may also find they are not learning useful material for their businesses and give poor satisfaction ratings for trainers. This would be particularly true for trainers whose incentives were based on 20 questions. Trainers, whose incentives were based on their student's answers to 80 questions, may find it best to teach well and impart useful knowledge, rather than focus on the answers to specific questions. This is because it may be that the best strategy to impart the knowledge on 80 topics is to provide broader-based knowledge, rather than attempting to cram 80 "facts" into a student's head in a short period.

We now examine the determinants of trainer performance as measured by satisfaction ratings. The results are presented in Table 6. Ordinary least squares would mis-estimate the standard errors on the group level variables if, as is likely, the error terms were correlated across groups. The estimation method we use is regression with the cluster command in STATA. This method controls for correlated error terms within groups and calculates robust standard errors. As student's ability may affect their enjoyment of the course, all regressions control for the student's pretest score. The variable is generally insignificant. Omitting it does not substantially change the results.

The first regression in Table 6 (Model 1) is a simple regression of the satisfaction ratings on whether or not the trainer had incentives. The coefficient is positive but insignificant, allowing no clear inference on the effect of incentives on the quality of teaching. (This result remains valid for the more complete models, below). This result is not surprising given the frequently aired concerns that giving incentives based on multiple choice questions would simply induce trainers to "teach for the test". The students—the SME managers who had come to take the course to improve their business practices—may, therefore, be dissatisfied with with such "narrow" instruction. We had anticipated this issue and planned to test for this concern. Our strategy was to offer incentives on outcomes of different breadths and we, therefore, gave different trainers incentives based on 20, 40 or 80 questions (see Table 4). This allows us to test whether incentives based on 80 questions are more motivating than incentives based on 20 questions.[20]

[20] Note, however, that the correlations between multiple choice scores and essay questions are similar for all three groups (20 questions, 40 questions and 80 questions).

We do this test of the effectiveness of incentives based on different numbers of questions in Models 2 through 4. These models include a dummy variable for the presence or absence of trainer incentives, the number of questions, the trainer quality variable, and their interaction terms. All variables are "standardized" in a way that allows an easy interpretation of the coefficients, with the exception of the variable, Questions, which has a value of -1 when there are 20 questions, 0 when there are 40 questions, and 2 when there are 80 questions.[21]

Table 6: The determinants of trainer performance, as measured by satisfaction ratings[2]

Dependant variable: Composite satisfaction ratings variable (Mean= 24.62, Standard deviation= 3.20)				
Name of explanatory variable	*Model*			
	1	2	3	4
	-0.122 1.28	-0.098 1.16	-0.114 1.37	-0.121 1.32
Incentives	0.819 1.00	0.321 0.79	0.156 0.40	0.139 0.31
(Questions-40)/20		-1.288** 8.77	-1.170** 8.14	-1.096** 6.74
Trainer score		0.281* 2.23	0.585** 4.82	0.343* 3.46
Incentives*Questions		0.586* 2.25	0.546+ 1.99	0.315 1.10
Trainer score*Questions			-0.115 1.69	-0.077 1.23
Incentives*Trainer score			-0.549* 2.10	-0.370 1.40
Incentives*Trainer score*Questions			0.350* 2.37	0.336* 2.34
Participant controls[1]	No	No	No	Yes
Number of observations	172	172	172	144
Adjusted R^2	0.01	0.13	0.14	0.13

Estimation method: Regression with standard errors adjusted for within group correlations. Absolute value of robust t statistics below coefficients.

+ significant at 10%; * significant at 5%; ** significant at 1%

[1] These include the participant's pre-test score, gender, whether the firm owner, number of employees at participant's firm, and whether participant's firm exports. Other variables like age were also insignificant but led to a large reduction in the number of observations.

[21] Note that this is not the standard way of standardization with mean 0 and standard deviation of 1, which in this instance would not have led to variables in the most easily interpretable form.

[2] Variables are standardized to help interpretations of interactions, with the exception of the variable, Questions, which has a value of -1 at 20 questions, 0 at 40 questions, and 2 at 80 questions. The interaction terms are composed of these standardized variables but has *not* itself been standardized.

Here, we see that the number of questions has a clear negative effect. One possible interpretation of this is that students do not like answering 80 questions and let their frustrations out when giving their satisfaction ratings of the trainer. Since donors or policy makers are typically not concerned about these short-term frustrations, if this was the reason for the question effect, then it can be discounted. We are going to proceed by making this "frustration" assumption and will turn to other interpretations later.

The term of interest in these models is the interaction term of incentives and the number of questions. This term is positive and significant, suggesting that incentives based on more questions are in fact more effective at raising student's satisfaction ratings than incentives based on fewer questions (t=2.25, P=0.033). The standardization allows us to easily interpret the coefficient on the interaction term. At 40 questions, the impact of incentives is simply the coefficient on incentives (an insignificant 0.32). At 20 questions it is 0.32-0.59=-0.27 (also insignificant). At 80 questions the impact is 0.32+2*0.59=1.50 (significant P=0.001).

We also ran another regression of satisfaction ratings on incentives separately for groups which answered 20, 40 and 80 questions. There was a negative effect when there were 20 questions, significant at the 7-percent level. A similar test showed that incentives had no effect on performance when there were 40 questions. For groups with 80 questions, incentives have a positive effect on satisfaction ratings, but it was only significant at the 13-percent level.

Another dimension to worry about when designing incentives is the existing quality of service. Some students of incentives believe that incentives encourage mediocrity. In dysfunctional schools, where teachers are routinely absent, incentives may be likely to improve performance. However, if performance is already excellent, then incentives based on easily measured outcomes are liable to change behavior without improving it. In other words, "If it ain't broke, don't fix it".

In fact, incentives of varying degrees of breadth and complexity may be appropriate for trainers of different ability. We know from our experience in teaching that we can more effectively motivate students of different ability by tests of varying levels of difficulty. The literature on management and organizational psychology also suggests that incentives based on outcomes

considered too hard by agents disheartens the agents, rather than motivating them (Bandura 1986; Locke and Latham 1990).

To test this hypothesis, we use Model 3 with all interactions involving incentives, the number of questions and trainer quality. This is the benchmark model we use.[22] The results are interesting. The base incentive term has no effect as before. The number of questions has a negative effect, which we interpret as a "frustration effect" as discussed above. Trainer score has a positive effect as expected. The interaction of incentives and the number of questions has a positive effect, which suggests that the multi-tasking concerns are real. As in Model 2, this term indicates that incentives based on broadly defined outcomes are most effective at motivating agents. The interaction of the trainer score and incentives has a negative effect suggesting that incentives are more appropriate for poorer trainers. The triple interaction term has a positive coefficient suggesting that incentives should be based on broader, more complex, outcomes for better trainers.

To interpret the coefficients as the effect of a treatment variable on outcomes we have to consider the entire set of interactions. We present these results in Table 7.[23] We find that incentives do matter significantly for average or above-average trainers when based on 80 questions (the P values for average and good trainers are 0.013 and 0.011 respectively). As suggested by the theoretical concerns about incentives based on narrow outcomes, incentives based on 20 questions appear to *actually worsen* the performance of good trainers. However, this effect is only significant at the 8-percent level. Incentives generally seem to improve the performance of poor trainers. This effect is clearest at 40 questions (largely because the standard errors are lowest for this group), but of similar magnitude for all three groups. All significant effects are close to or larger than 1, and thus have meaningfully large magnitudes, close to or larger than one-third of the standard deviation of the dependant variable.

[22] Model 4 is the same but with a series of participant control variables. See discussion below.

[23] The magnitude calculated is the coefficient on incentives,
+ the question variable*the coefficient on the interaction of incentives and the question variable
+ the trainer quality variable*the coefficient on the interaction of incentives and trainer quality
+ the trainer quality variable*the question variable*the coefficient on the interaction of incentives, and trainer quality and the question variable.
The P-value is calculated using the "test" command in STATA which calculates the F-stat and the P value for linear combinations of coefficients

Model 4 is the same as Model 3 but with a series of participant control variables, including gender, whether owner of firm, number of employees of firm, and whether firm is an exporter. Other variables like participant age were also considered. The significance level on the triple interaction term is unchanged when participant controls are added. However these control variables are singly and jointly insignificant.[24] Hence, we focus on Model 3 without participant controls as our benchmark.

Table 7: The impact of incentives on outcomes.

Number of questions	*Poor trainers*	*Average trainers*	*Good trainers*
20	1.22 0.098	-0.390 0.511	-2.00 0.071
40	1.14 0.028	0.155 0.693	-0.832 0.252
80	0.97 0.1284	1.247 0.013	1.519 0.011

Notes: P-values below implied impact.

How can we interpret these results? It appears that for average or above average trainers, incentives do improve performance but only if based on 80 questions—and, perhaps, 40 questions for poor trainers. However, for trainers of below average performance (in the pre-training assessment) incentives appear to improve performance. This corresponds with the modern view of incentives, that if performance is reasonably good, incentives can be useful, but only if based on a broad measure of success. If initial performance is poor however, incentives appear to improve performance.

As always, some caveats bear making. For example, it is important to remember that if the negative coefficient on the number of questions in the regressions is interpreted differently, for instance, as "teachers teach badly if they feel they have to cover too much material" then the results presented here would have to be reinterpreted. Incentives according to this alternative interpretation may not be appropriate. For these and other reasons, we hope to have the opportunity to conduct another study where we can sharpen our methodology with the benefit of hindsight, and get clearer results in the future. The next section covers the lessons we have learned from this PREP application that will help us refine future such studies.

[24] The controls' effects appear to be neutralized by the ensuing reduction of the number of observations, since the adjusted R^2 actually declines with them included.

5 Relaxation of PREP requirements[25]

Section 3 provided a list of requirements for conducting a successful prospective, randomized evaluation. These were illustrated in the SME application above. However, it may be the case, especially when implementing a PREP for an institutional reform, that the "experimenter" is unable to obtain full control of the treatment to be evaluated so that these conditions cannot all be met. A simple example is that some jurisdictions may not want to adopt the recommended reforms. In this section, we consider the scope for running PREP under weaker sets of experimental conditions.

The technique we propose is frequently used by economists in dealing with real-world data: two-stage least squares (2SLS) or instrumental variable (IV) estimation.[26] There are several reasons to be skeptical of IV estimation when applied to real-world data, but we will show that these concerns are not serious when applied to a quasi-randomized implementation that satisfies the conditions described below. First the reader may wish to review the ideal conditions for precise inference under a PREP application, as described in Table 3.

An instrumental variable technique allows the relaxation of Condition 5 (treatment group adopts the reform just as they would if the reform were broadly implemented) and Condition 6 (control group continues to act as if the reform were not implemented) but makes more strenuous demands on Condition 3 (reform is implemented according to a clear randomization protocol) and 4 (there must be a sufficiently large number of "individuals" in the control and treatment groups).

We should emphasize that maintaining Conditions 5 and 6 can be extremely difficult in real world situations. It will often be the case that some municipalities in the treatment group will not adopt certain reforms. Similarly, municipalities in the control group may adopt reforms. The first concern is more likely for popular reforms and the second one for unpopular reforms.

It is worth addressing the question of how a violation of Conditions 5 and 6 would affect a simple comparison of the treatment and control groups. If, for some exogenous reason uncorrelated with outcomes, a fraction of the treatment group failed to undertake the reform and a fraction of the control group adopted the reform, then the impact of the reform would be under-

[25] We are grateful to Peter Murrell for suggesting that we discuss how the conditions for a randomized evaluation could be relaxed.

[26] These two methods are very similar. The reader interested in a description of these methods should refer to Davidson and McKinnon (1993).

estimated. The magnitude of the bias could be quite substantial. For instance, if 25 percent of the control group undertook the reform and 25 percent of the treatment group did not, then the impact of the reform would be underestimated by 50 percent. Since the magnitude of the impact of policy reforms is as important as whether or not they had an effect at all, such an under-estimate could have important and unfortunate consequences such as under-funding of reform.

In fact, the adoption or non-adoption of reform may not be random. The most reform-minded of the control group might adopt the reform, and reform may be most effective in these places where they are most sincerely adopted. In this case the bias would be even larger.

Let us proceed with an example. Consider the effects of an improvement in procurement practices on the level of corruption. Municipalities may have been instructed to advertise procurements more widely, receive more competitive bids, and document the basis on which they make their decisions. Say the reform was consciously implemented in 100 municipalities and not in another 100. However, a few of the municipalities in the treatment group did not adopt all the reforms and some of the municipalities in the control group adopted some. How is the PREP evaluator to proceed in the anticipation this happening?

The PREP evaluator should anticipate some leakage in one direction or the other and collect information on variables that are likely to affect the propensity to adopt reforms. Examples of variables to use might include indicators of corruption (especially of the state capture kind) and general reform-mindedness of the municipal government. A pre-survey and focus group discussions can help identify in this regard.

The second step is to measure the degree of reform implementation. The PREP evaluator should create a methodology to measure implementation on a scale from 0 to 1 in all municipalities in the control *and* treatment groups, where 0 refers to "no reforms adopted" and 1 to "all reforms adopted".[27] In our current example, checking procurement reform implementation would mean checking that procurement bids were being advertised more widely, competitive bids were being received, and the basis for procurement decisions was being properly documented. These checks are *not* a substitute for measuring outcomes. Rather, they are an additional step that must be taken more seriously if IV estimation methods are to be used. The next step is to complete the measurement of outcomes etc. as would be done in a standard PREP.

[27] While the degree of reform implementation should be checked in any case, if IV estimation methods are not used then simply inspecting a few randomly selected municipalities might be enough.

Finally, when it comes to analysis, whether a municipality was in the treatment group must be used as an instrument for adopting reform. Essentially this means using whether the municipality was in the treatment group to predict whether reform was adopted, and then correlating this predicted value with improvements in performance. Both models should be estimated using reform-mindedness as a control because the omission of these variables is likely to lead to biased results. As long as the degree of reform implementation and the variables which predict adoption are adequately measured, the impact of reform can be measured with reasonable accuracy.

To state the above formally, the following model must be estimated:

$$\text{Outcome} = ß_0 + ß_1 \text{ Degree implemented} + ß_2 \text{ Reform-minded} + e$$

$$\text{Degree implemented} = ?_0 + ?_1 \text{ Treatment group dummy} + ?_2 \text{ Reform-minded} + ?$$

We now turn to the issue of the validity of the assumptions of IV estimation in this quasi-randomized case. To put it simply, when examining the impact of *A* on *B*, an instrument *C* must satisfy the following conditions.

1. *C* affects *A*
2. *C* does not directly affect *B*
3. There is no *D* that might directly affect both *C* and *B*.

It is difficult to find such a variable in real-world data. However, a randomization protocol clearly satisfies Conditions 2 and 3 (as long as the assignment of those receiving treatment is truly random and not subject to political manipulation). The first criterion must be ensured by trying earnestly to implement reforms in the treatment group but not the control group. While the IV method can deal with some leakages, too much leakage would lead to violation of the first condition and make the IV estimator inaccurate.

6 Lessons learned

Oscar Wilde once quipped "experience is the name we give to our mistakes" and indeed we have learnt much from ours.

The first predictable lesson is "collect more data". While this is always true, we seem to have erred on the side of collecting too little data in the sense that the marginal usefulness of

more data would have quite high. We did in fact try to collect more data. We had specified a larger number of participants than eventually appeared on the day of the training (450 rather than 274). Our subcontractor tried hard with phone calls and advertisements in newspapers and even television to increase enrollments but this is the best they could manage. Budgetary and time restrictions prevented our conducting another set of trainings in Cebu and other major cities, but perhaps we should have fought for more time and resources.

We also erred by assuming that there would be less individual-level "noise" than there actually was in all the variables other than the multiple choice scores. Normally pre-tests may have alleviated this risk by helping to definitively establish the necessary sample size. The presence of these individual-level factors, together with the small number of observations in each classroom, made all the performance variables, other than satisfaction ratings, unreliable indicators of the quality of teaching. Had we been able to anticipate this problem—say, from a pre-test, we would have conducted one or two trainings where the entire population of students was taught by the same trainer. Responses on the willingness to pay for more trainings from this trainer, satisfaction ratings of this trainer's performance, and statements about whether these trainings would affect the way business would be conducted could be used to "clean out" individual-level effects. However, we should not overstate the expected effectiveness of such a method. For example, we did anticipate this problem for the "test-scores" variable and conducted a pre-test. Cleaning the post-training score using the pre-test score did not significantly improve the performance of the test-score variable.

It is also a good idea to collect data on an intermediate performance indicator when conducting a PREP. As discussed in Section 5, this allows the evaluators to deal with leakages from the treatment to the control group and vice-versa.

Another lesson relates to the appropriate incentives to motivate trainers. We found that some of the trainers were motivated by the desire to work for ISSI or IRIS in the future—even though we had explicitly stated this was a one-off IRIS project, and there were no future plans for collaboration with ISSI. Such motivations could easily drown out the monetary incentives we had provided. Several possible solutions could be considered to remedy this problem. First, one could provide much larger financial incentives. Second, one could employ an experimental design that is semi-blind so as to prevent (in the present case) IRIS and ISSI from learning the trainers' true performance. Finally, one could require the subcontractor to sign an agreement not

to hire any of the trainers for say three years. In the latter two cases, the strategy is to convince the trainers *ex ante* that IRIS and ISSI have credibly pre-committed to having no future plans for collaboration with them.

We should perhaps also have had a group of trainers rewarded on the basis of scores on essay questions and multiple choice tests. This would have substantially broadened the basis for incentives. A comparison between this group and the other groups would have been instructive. Likewise, the essay question should have been temporally administered randomly, either before or after the multiple-choice questions to control for the "fatigue" effect.

In terms of policy advice on offering incentives, our results suggest that only incentives based on broadly defined outcomes are likely to work. This corresponds with theoretical conjectures, concerns expressed in the education literature, and concurrent empirical work. However, we must end on a note of calling for more research on this subject before definitive policy-reform conclusions can be drawn. More studies structured with the improvements we have suggested here are needed in more countries to develop a clear understanding on how explicit monetary incentives motivate teachers and trainers.

References

Azfar, O. (2002), "The NIE approach to economic development: An analytical primer", *IRIS Discussion Papers on Institutions and Development*, College Park, Maryland, USAID SEGIR/LIR Task Order 7.

Azfar, O., S. Kähkönen, and P. Meagher (2001) "Conditions for Effective Decentralized Governance: A Synthesis of Research Findings", *mimeo*, IRIS Center, University of Maryland.

Azfar, O., P. Murrell, with contributions from A. Lanyi and M. Russell-Einhorn (2003), "Assessing Corruption and Institutional Integrity: A Handbook Describing a Profile of Institutional Integrity–Measurement and Assessment Toolkit (PII-MAT)" *mimeo*, The IRIS Center of the University of Maryland.

Bandura, A. (1986), Social foundations of thought and action: A social cognitive theory, *Englewood Cliffs, NJ Prentice Hall.*

Davidson and Mckinnon (1993), Estimation and inference in econometrics, Oxford University Press**.**

Dixit, A. (2001), "Incentives and organizations in the public sector: An interpretative review", *mimeo*, Princeton University.

Ebert, R., K. Holleneck and J. Stone (2002), "Teacher performance incentives and student outcomes", *mimeo*, Department of Economics, University of Oregon.

Glewwe, P., N. Ilias and M. Kremer, Teacher incentives, *mimeo Harvard 2003.*

Holmstrom, B. and P. Milgrom (1991), "Multi-task Principal Agent Analysis: Incentive Contracts, Asset Ownership and Job Design", *Journal of Law, Economics and Organization* 7(0), 24-52.

Klitgaard, R. (1988). *Controlling Corruption*, Berkeley: University of California Press.

Klitgaard, R., R. Maclean-Abarora and H. L. Parris (2001), *Corrupt cities: A practical guide to cure and prevention* ICS Press.

Kremer, M. (2002), "Innovations for Donor Assistance", *IRIS Discussion Papers on Institutions and Development*, College Park, Maryland, F1-3, USAID SEGIR/LIR Task Order 7.

Kremer, M. (2003) "Randomized evaluations of educational programs in developing countries: Some lessons" *forthcoming AER papers and proceedings, May 2003.*

Lederman, D., N. Loyaza and R. Soares (2002), "Accountability and corruption: Political Institutions Matter", *mimeo*, The World Bank.

Lee, Y. and O. Azfar (2000). "Does Corruption Delay Trade Reform?" *IRIS Working Papers*, The IRIS Center, University of Maryland.

Locke, E.A. and G.P. Latham, (1990), *A theory of goal setting and task performance*, Englewood Cliffs, NJ: Prentice Hall.

Murnane, R. and D. Cohen (1986), "Merit pay and the valuation problem: Why most merit pay plans fail and a few survive", *Harvard Educational Review*, 56-1, 1-17.

Prendergast, C. (1999), "The provision of incentives in firms", *Journal of Economic Literature*, Vol 37, 7-6].

Rauch, J. and P. Evans (2000), "Bureaucratic structure and bureaucratic performance in less developed countries", *Journal of Public Economics*, **75**, 49-71.

Van Rijckeghem, C. and B. Weder (1997). "Corruption and the Rate of Temptation: Do Low Wages in the Civil Service Cause Corruption?," *IMF Working Paper*, June, WP/97/73.

Zinnes, C. and O. Azfar (2003), "Transactions Cost Analysis and Remediation," *IRIS Discussion Papers on Institutions and Development*, College Park, Maryland, F5-3, USAID SEGIR/LIR Task Order 7.

Appendix A

Table A.1 Means of performance variables

Variable	Observation	Mean	Std. Dev.	Min	Max
G3	262	3.59	0.563	1	4
G4	259	3.65	0.523	2	4
G5	191	3.61	0.548	2	4
G6	260	3.42	0.638	1	4
G7	257	3.40	0.649	1	4
G8	260	3.38	0.643	2	4
G13	257	3.48	0.593	1	4
H3	243	1.06	0.255	1	2
H5	142	1.47	0.501	1	2
H7	251	1.61	0.986	1	4
H8	255	1.39	0.829	1	4
H9	249	1.55	0.949	1	4
H10	253	1.54	0.948	1	4
H11	246	1.93	1.143	1	4
H12	206	2.17	1.109	1	4
H13	232	1.75	1.067	1	4
H14	212	2.03	1.122	1	4
H15	229	1.48	0.962	1	5
H16	184	1.46	0.957	1	5
H17	215	1.57	0.943	1	5
H18	183	1.55	0.940	1	5

Appendix B: Performance indicator questions

B.1: Student trainer-satisfaction rating questions

Please rank on a scale of 1-4

G3. Could you understand the lecture

G4. Did the lecturer speak.. 1=can hardly be heard 4=loud

G5. Did the lecturer speak 1=difficult…4=easy

G6. How well did the lecturer present his/her topics

G7. How well did the lecturer conduct class discussion

G8. How well did the lecturer satisfy inquiries from participants

G13What is your overall rating of the lecture

B.2: Questions to elicit the effect of training on the way participants engage in marketing

H3. Will the training you just received help you increase your exports?

H5. Will the training you just received help you increase your sales?

H7. Will the training you just received help you find market niches?

H8. Will the training you just received help you improve your product?

H9. Will the training you just received help you change the way you provide after sales service?

H10. Will the training you just received change the way you conduct product enhancement?

B.3: Willingness-to-pay questions for subsequent trainings on different subjects by the same trainer

Normally ISSI would charge P1000-P2000 for a one day training. Would you pay P 1000 for the following topics?

H11. Preparing an Effective Strategic Marketing Plan

H12. Gearing up for AFTA (Asian Free Trade Agreement)

H13. Sales Forecasting Techniques

H14. Export Marketing

Appendix C: Trainer incentive payment schedule

1$=approximately 50 pesos
Per-Capita GDP = 45,490 Pesos in 2001. (Trainers can reasonably be assumed to be in top decile of income earners. In 1997 the top decile earned twice as much as the country average).

For 6 trainers:

Dear Trainer (PERSONALIZE)

We have determined on the basis of your performance in the training on the 10th of November and after examining your resume to offer you some additional compensation. PLEASE DO NOT DISCUSS THIS ADDITIONAL COMPENSATION WITH ANYONE. We have tried to ensure that all trainers will get the same additional compensation in expected terms.

You will be given 3000 Pesos in addition to the $400 that was agreed between you and ISSI. Of course all compensation will be taxed.

For 18 trainers (3 groups of 6). Some get 20 questions, some 40, and some 80:

Dear Trainer (PERSONALIZE)

We have determined on the basis of your performance in the training on the 10th of November and after examining your resume to offer you some additional compensation. PLEASE DO NOT DISCUSS THIS ADDITIONAL COMPENSATION WITH ANYONE. We have tried to ensure that all trainers will get the same additional compensation in expected terms.

In addition to the $400 you agreed with ISSI your compensation will depend on the average performance of your class on the enclosed multiple choice questions according to the table below. Of course all compensation will be taxed.

Class performance	Your additional compensation
50% or below	0
51-55%	1000
56-60%	2000
61-65%	3000
66-70%	4000
71-75%	5000
76-80%	6000
81-85%	7000
86-90%	8000
91-95%	9000
96-100%	10000